Finetuning Llama 2 7b Foundational Model for Healthcare Dataset

1. Pre-trained Model Evaluation

1.1. Deploy the Llama2 Model on AWS Sagemaker

The next cell will take some time to run. It is deploying a large language model, and that takes time. You'll see dashes (--) while it is being deployed. Please be patient! You'll see an exclamation point at the end of the dashes (---!) when the model is deployed and then you can continue running the next cells.

You might see a warning "For forward compatibility, pin to model_version..." You can ignore this warning, just wait for the model to deploy.

[4]:
```
from sagemaker.jumpstart.model import JumpStartModel

model = JumpStartModel(model_id=model_id, model_version=model_version, instance_type="ml.g5.2xlarge")
predictor = model.deploy()
```

```
For forward compatibility, pin to model_version='2.*' in your JumpStartModel or JumpStartEstimator definitions. Note that major version
upgrades may have different EULA acceptance terms and input/output signatures.
Using vulnerable JumpStart model 'meta-textgeneration-llama-2-7b' and version '2.1.8'.
Using model 'meta-textgeneration-llama-2-7b' with wildcard version identifier '2.*'. You can pin to version '2.1.8' for more stable resu
lts. Note that models may have different input/output signatures after a major version upgrade.
------------------!
```

1.2. Evaluate the Pre-trained Llama2 Text Generation Large Language Model for Domain Knowledge

jupyter Model_Evaluation_UdacityGenAIAWS Last Checkpoint: 13 minutes ago (unsaved changes) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted conda_pytorch_p310

Run Markdown nbdiff

In [6]:
```
payload = {
    "inputs": "Myeloid neoplasms and acute leukemias derive from",
    "parameters": {
        "max_new_tokens": 64,
        "top_p": 0.9,
        "temperature": 0.6,
        "return_full_text": False,
    },
}
try:
    response = predictor.predict(payload, custom_attributes="accept_eula=true")
    print_response(payload, response)
except Exception as e:
    print(e)
```

```
Myeloid neoplasms and acute leukemias derive from
>  myeloid stem cells and are characterized by the presence of immature myeloid cells, including promyelocytes, myelocytes, metamyelocytes, and band cells. M
yeloid neoplasms are classified as acute myeloid leukemia (AML)

==================================
```

In [7]:
```
payload = {"inputs": "Genomic characterization is essential for",
           "parameters": {"max_new_tokens": 64, "top_p": 0.9, "temperature": 0.6, "return_full_text": False,},}
try:
    response = predictor.predict(payload, custom_attributes="accept_eula=true")
    print_response(payload, response)
except Exception as e:
    print(e)
```

```
Genomic characterization is essential for
>  the development of molecular diagnostics.
The authors thank the members of the UMMS-MGH Center for Molecular Discovery for their assistance in the development of the method. The authors also thank th
e members of the UMMS-MGH Center for Molecular Discovery for their

==================================
```

In [8]:
```
payload = {"inputs": "Certain germline disorders may be associated with",
           "parameters": {"max_new_tokens": 64, "top_p": 0.9, "temperature": 0.6, "return_full_text": False,},}
try:
    response = predictor.predict(payload, custom_attributes="accept_eula=true")
    print_response(payload, response)
except Exception as e:
    print(e)
```

jupyter Model_Evaluation_UdacityGenAIAWS Last Checkpoint: 15 minutes ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted conda_pytorch_p310

Run Markdown nbdiff

In [8]:
```
payload = {"inputs": "Certain germline disorders may be associated with",
           "parameters": {"max_new_tokens": 64, "top_p": 0.9, "temperature": 0.6, "return_full_text": False,},}
try:
    response = predictor.predict(payload, custom_attributes="accept_eula=true")
    print_response(payload, response)
except Exception as e:
    print(e)
```

```
Certain germline disorders may be associated with
>  an increased risk of developing malignant neoplasms. The presence of a germline mutation in a gene associated with cancer predisposition may influence the
choice of therapy, and the timing of surveillance for second neoplasms. The aim of this study was to evaluate the preval

==================================
```

In [9]:
```
payload = {"inputs": "In contrast to targeted approaches, genome-wide sequencing",
           "parameters": {"max_new_tokens": 64, "top_p": 0.9, "temperature": 0.6, "return_full_text": False,},}
try:
    response = predictor.predict(payload, custom_attributes="accept_eula=true")
    print_response(payload, response)
except Exception as e:
    print(e)
```

```
In contrast to targeted approaches, genome-wide sequencing
>  has the advantage of identifying variants of unknown significance.
We performed whole-genome sequencing on 1,372 individuals, including 1,298 individuals with autism and 74 controls. We found that autism is associated with r
are genetic variants and that the spectrum of gen

==================================
```

The prompt is related to the domain you want to fine-tune your model on. You will see the outputs from the model without fine-tuning are limited in providing insightful or relevant content.

Use the output from this notebook to fill out the "model evaluation" section of the project documentation report

Take a screenshot of this file with the cell output for your project documentation report. Download it with cell output by making sure you used Save on the notebook before downloading

After you've filled out the report, run the cells below to delete the model deployment

`IF YOU FAIL TO RUN THE CELLS BELOW YOU WILL RUN OUT OF BUDGET TO COMPLETE THE PROJECT`

2. Fine-tuning a Large Language Model

2.1. Fine-tune a Large Language Model with a Domain-Specific Dataset

```
[4]: from sagemaker.jumpstart.estimator import JumpStartEstimator
     import boto3

     estimator = JumpStartEstimator(model_id=model_id,  environment={"accept_eula": "true"},instance_type = "ml.g5.2xlarge")

     estimator.set_hyperparameters(instruction_tuned="False", epoch="5")

     #Fill in the code below with the dataset you want to use from above
     #example: estimator.fit({"training": f"s3://genaiwithawsproject2024/training-datasets/finance"})
     # s3://genaiwithawsproject202406/training-datasets/medicalDataset.txt
     estimator.fit({"training": f"s3://genaiwithawsproject202406/training-datasets/medicalDataset.txt"})
```

```
PEFT modules are saved in saved_peft_model directory
best eval loss on epoch 4 is 2.4424901008605957
Epoch 5: train_perplexity=7.6446, train_epoch_loss=2.0340, epcoh time 9.125123106000046s
INFO:root:Key: avg_train_prep, Value: 7.934391975402832
INFO:root:Key: avg_train_loss, Value: 2.0708813667297363
INFO:root:Key: avg_eval_prep, Value: 11.927809715270996
INFO:root:Key: avg_eval_loss, Value: 2.4785332679748535
INFO:root:Key: avg_epoch_time, Value: 9.401695886800008
INFO:root:Key: avg_checkpoint_time, Value: 0.7507950958000151
INFO:root:Combining pre-trained base model with the PEFT adapter module.
Loading checkpoint shards:   0%|          | 0/2 [00:00<?, ?it/s]
Loading checkpoint shards:  50%|[redacted]| 1/2 [00:29<00:29, 29.77s/it]
Loading checkpoint shards: 100%|[redacted]| 2/2 [00:35<00:00, 15.48s/it]
Loading checkpoint shards: 100%|[redacted]| 2/2 [00:35<00:00, 17.63s/it]
INFO:root:Saving the combined model in safetensors format.
INFO:root:Saving complete.
INFO:root:Copying tokenizer to the output directory.
INFO:root:Putting inference code with the fine-tuned model directory.
2024-06-14 15:48:52,021 sagemaker-training-toolkit INFO     Waiting for the process to finish and give a return code.
2024-06-14 15:48:52,021 sagemaker-training-toolkit INFO     Done waiting for a return code. Received 0 from exiting process.
2024-06-14 15:48:52,022 sagemaker-training-toolkit INFO     Reporting training SUCCESS

2024-06-14 15:48:59 Uploading - Uploading generated training model
2024-06-14 15:49:42 Completed - Training job completed
Training seconds: 701
Billable seconds: 701
```

3. Evaluate the Fine-tuned Llama2 Large Language Model

3.1. Deploy the Fine-tuned Llama2 Model on AWS Sagemaker

Deploy the fine-tuned model

Next, we deploy the domain fine-tuned model. We will compare the performance of the fine-tuned and pre-trained model.

```
[5]: finetuned_predictor = estimator.deploy()
```

```
No instance type selected for inference hosting endpoint. Defaulting to ml.g5.2xlarge.
INFO:sagemaker.jumpstart:No instance type selected for inference hosting endpoint. Defaulting to ml.g5.2xlarge.
INFO:sagemaker:Creating model with name: meta-textgeneration-llama-2-7b-2024-06-14-15-51-19-257
INFO:sagemaker:Creating endpoint-config with name meta-textgeneration-llama-2-7b-2024-06-14-15-51-19-254
INFO:sagemaker:Creating endpoint with name meta-textgeneration-llama-2-7b-2024-06-14-15-51-19-254
-----------!
```

3.2. Evaluate the Fine-tuned Llama2 Text Generation Large Language Model on Text Generation Tasks and Domain Knowledge

[7]:
```
payload = {
    "inputs": "Myeloid neoplasms and acute leukemias derive from",
    "parameters": {
        "max_new_tokens": 64,
        "top_p": 0.9,
        "temperature": 0.6,
        "return_full_text": False,
    },
}
try:
    response = finetuned_predictor.predict(payload, custom_attributes="accept_eula=true")
    print_response(payload, response)
except Exception as e:
    print(e)
```

```
Myeloid neoplasms and acute leukemias derive from
> [{'generated_text': ' myeloid progenitor cells. Myeloid progenitor cells are present in the bone marrow, but they can also be found in
the peripheral blood, in the spleen, and in the liver. Myeloid progenitor cells are the precursors of white blood'}]

==================================
```

[8]:
```
payload = {"inputs": "Genomic characterization is essential for",
           "parameters": {"max_new_tokens": 64, "top_p": 0.9, "temperature": 0.6, "return_full_text": False,},}
try:
    response = finetuned_predictor.predict(payload, custom_attributes="accept_eula=true")
    print_response(payload, response)
except Exception as e:
    print(e)
```

```
Genomic characterization is essential for
> [{'generated_text': ' the identification of genetic variants that may affect drug response, but the cost of sequencing has limited the
application of this approach to clinical care. Genetic variants identified in the context of clinical trials can be used to inform clini
cal practice, but their clinical utility is limited by their rarity and'}]

==================================
```

[9]:
```
payload = {"inputs": "Certain germline disorders may be associated with",
           "parameters": {"max_new_tokens": 64, "top_p": 0.9, "temperature": 0.6, "return_full_text": False,},}
try:
    response = finetuned_predictor.predict(payload, custom_attributes="accept_eula=true")
    print_response(payload, response)
except Exception as e:
    print(e)
```

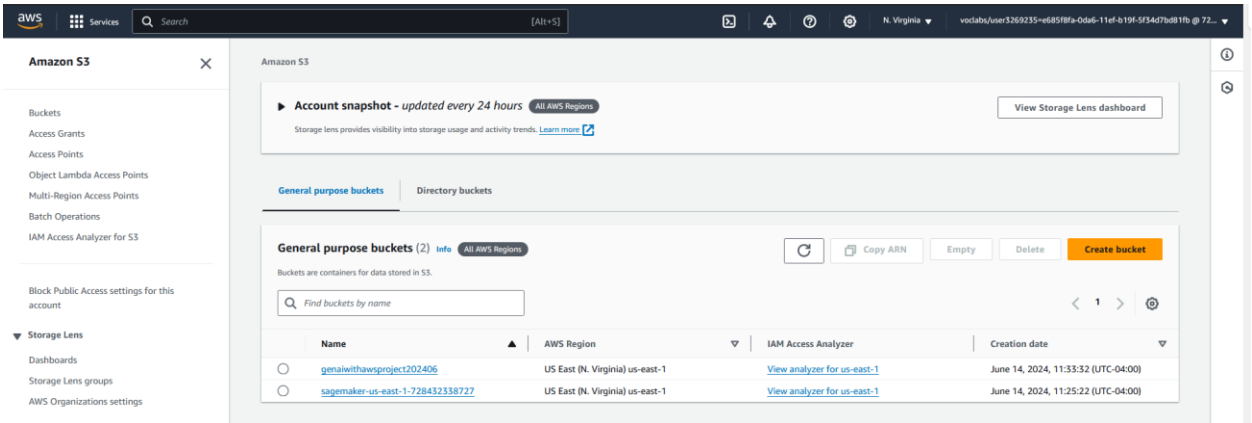
```
Certain germline disorders may be associated with
> [{'generated_text': ' an increased risk of developing certain types of cancer. Some of these disorders are inherited from a parent, wh
ile others may be acquired later in life.\nGermline genetic testing is a type of genetic testing that looks for changes in your DNA that
may increase your risk of developing cancer. This type of'}]

==================================
```

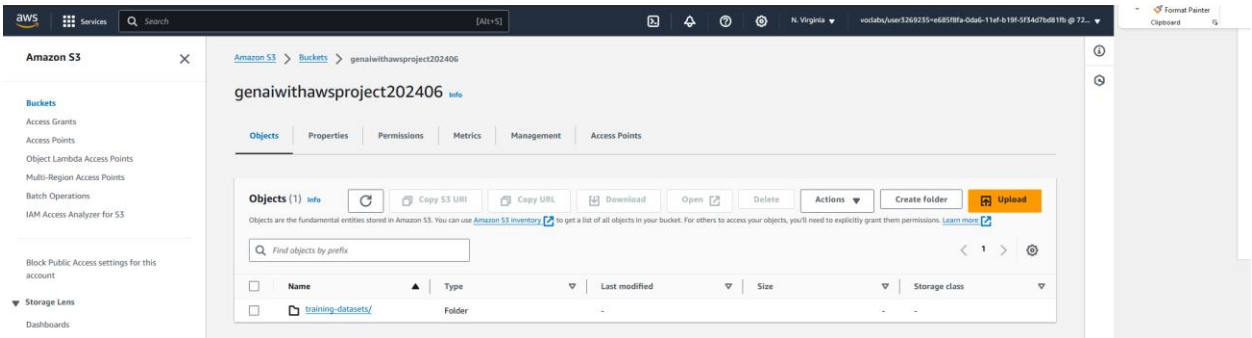
[10]:
```
payload = {"inputs": "In contrast to targeted approaches, genome-wide sequencing",
           "parameters": {"max_new_tokens": 64, "top_p": 0.9, "temperature": 0.6, "return_full_text": False,},}
try:
    response = finetuned_predictor.predict(payload, custom_attributes="accept_eula=true")
    print_response(payload, response)
except Exception as e:
    print(e)
```

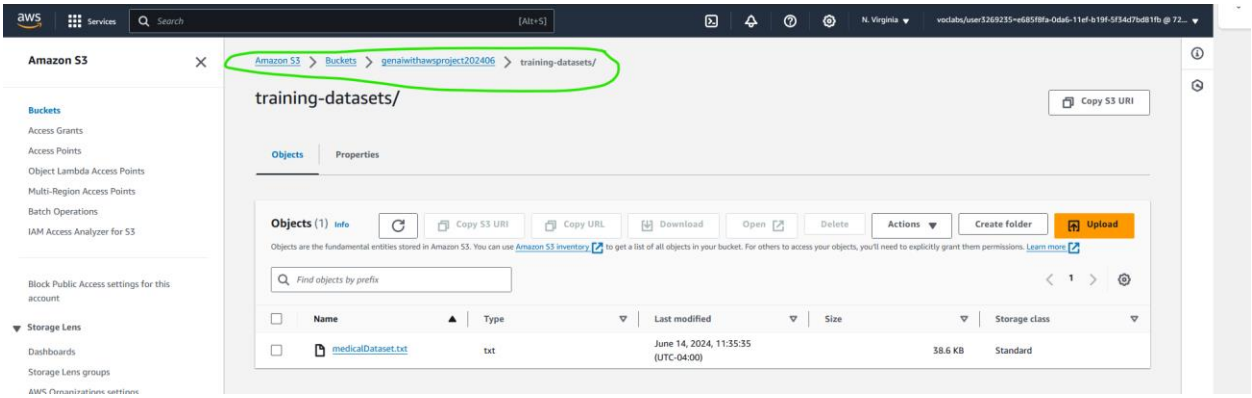
```
In contrast to targeted approaches, genome-wide sequencing
> [{'generated_text': ' of a large number of tumors from the same cancer type can reveal the genomic landscape of the disease and thus i
dentify potential drivers of the disease.\nThe aim of this study is to establish a large-scale sequencing of primary tumors and metastas
es from patients with colorectal cancer. The'}]

==================================
```

4. S3 Bucket screenshots

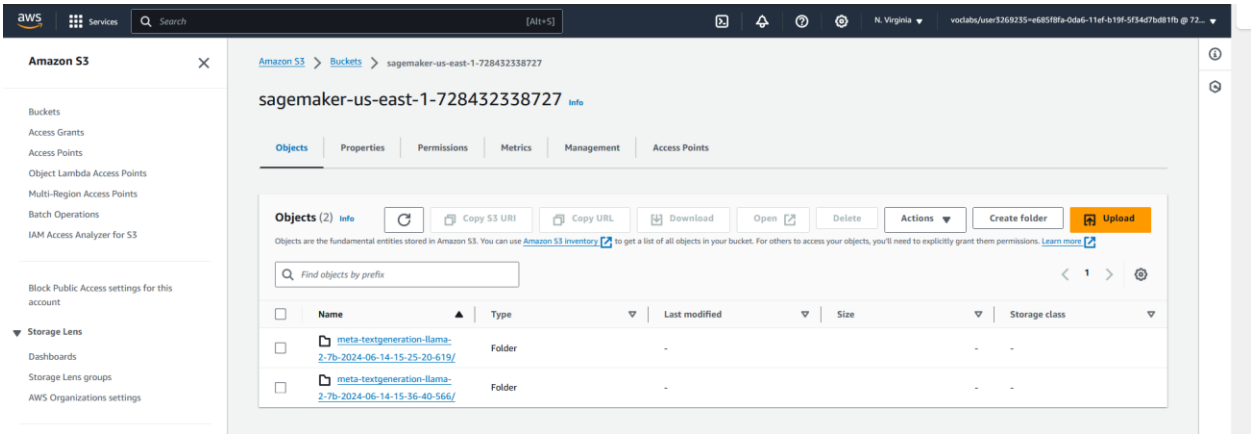
4.1. Healthcare dataset location

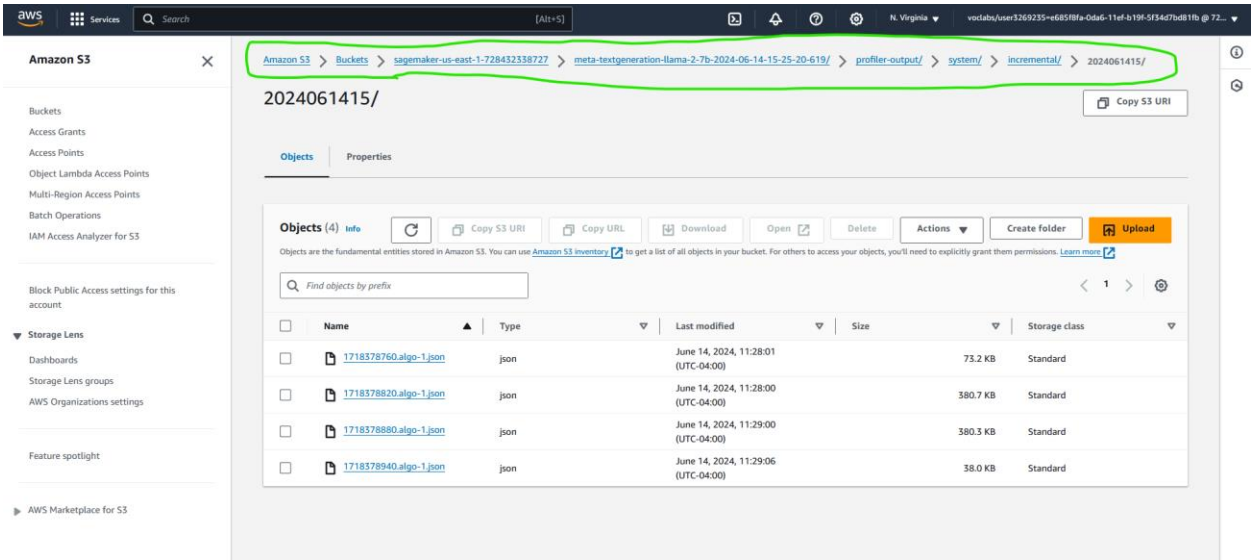


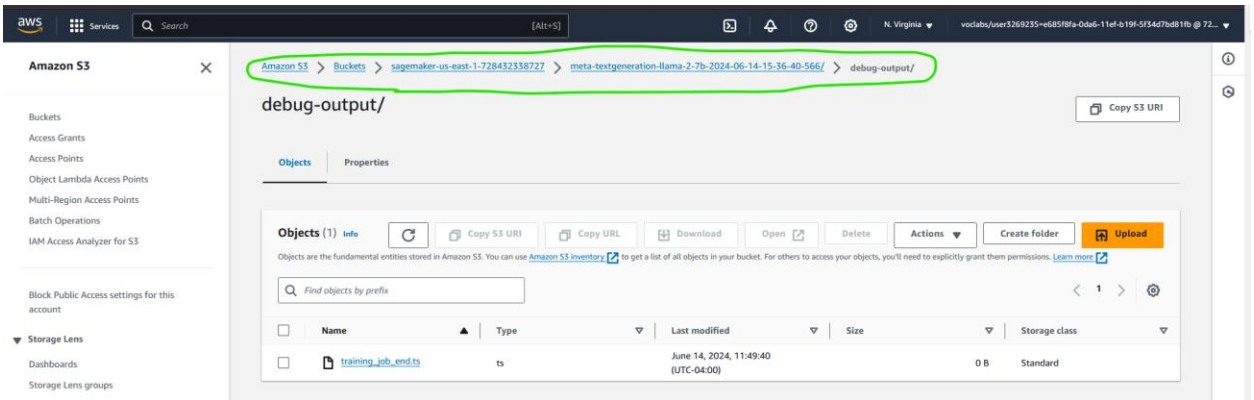


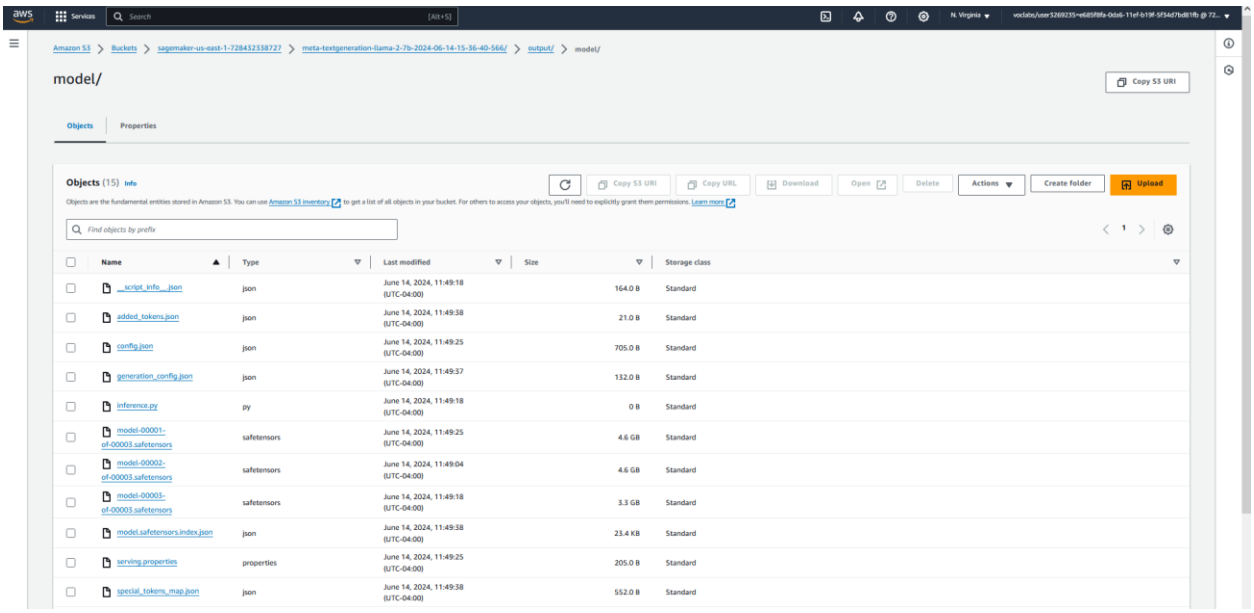


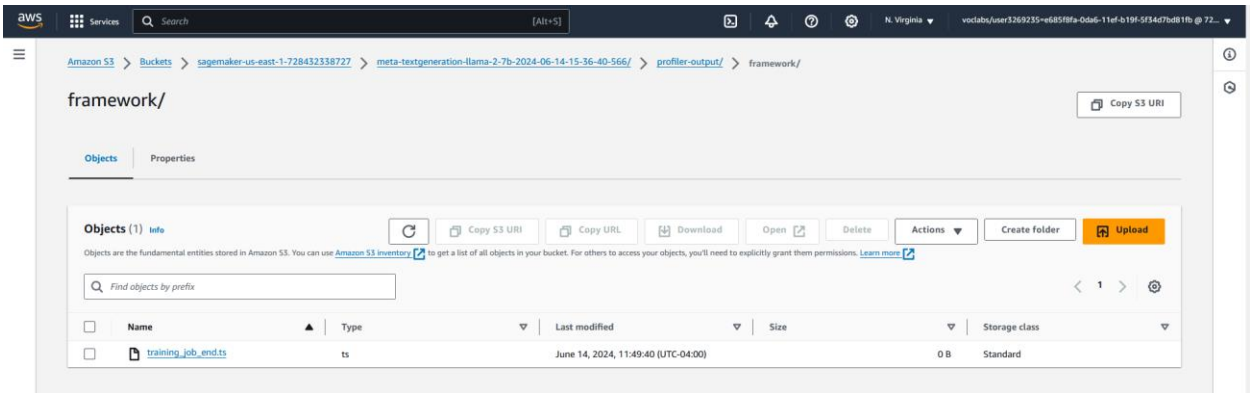
4.2. Metadata - fine-tuned model weights

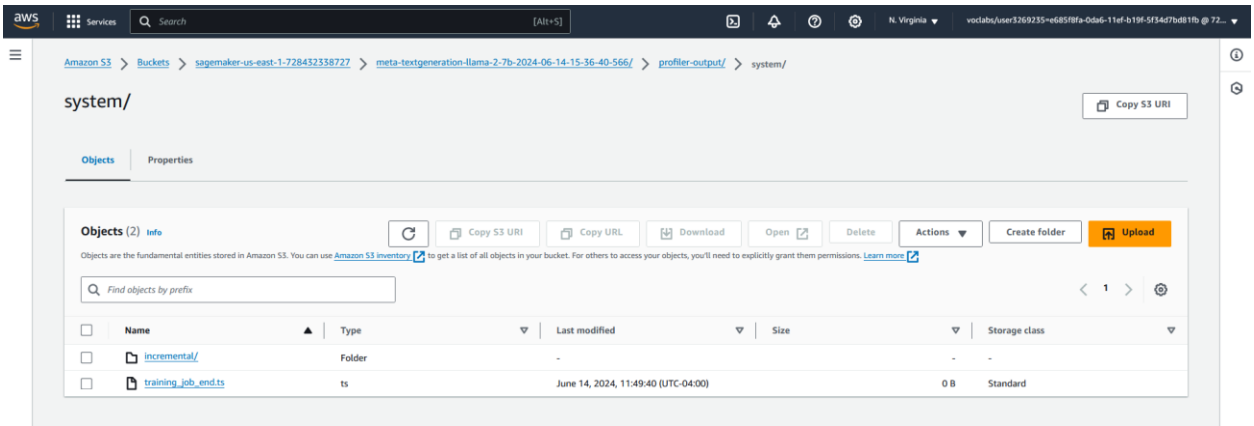


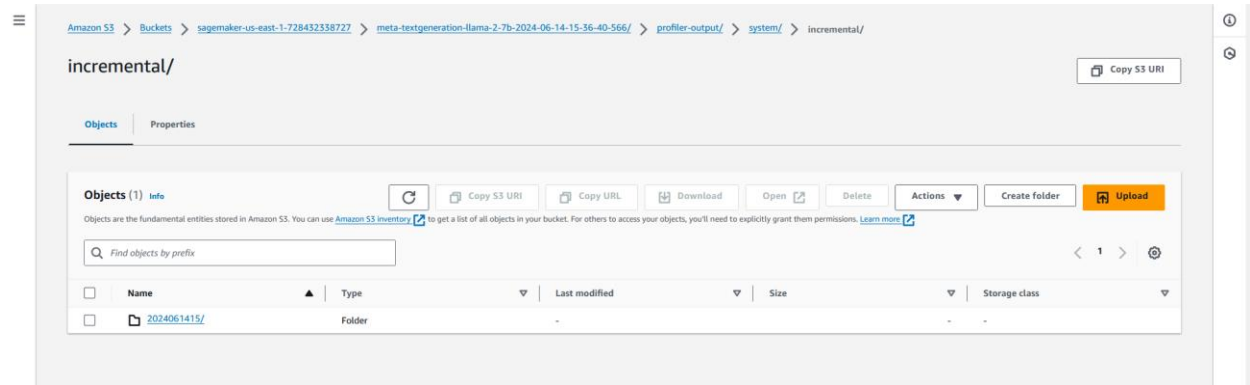



aws
Services
Search
[Alt+S]
N. Virginia
voclabs/user3269235=e685f8fa-0da6-11ef-b19f-5f54d7bd81fb @ 72...
Amazon S3
Buckets
Access Grants
Access Points
Object Lambda Access Points
Multi-Region Access Points
Batch Operations
IAM Access Analyzer for S3
Block Public Access settings for this account
Storage Lens
Dashboards
Storage Lens groups
Amazon S3 > Buckets > sagemaker-us-east-1-728432338727 > meta-textgeneration-llama-2-7b-2024-06-14-15-36-40-566/ > debug-output/
debug-output/
Copy S3 URI
Objects
Properties
Objects (1) Info
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more
Find objects by prefix
Name
Type
Last modified
Size
Storage class
training_job_end.ts
ts
June 14, 2024, 11:49:40 (UTC-04:00)
0 B
Standard


Amazon S3 > Buckets > sagemaker-us-east-1-728432338727 > meta-textgeneration-llama-2-7b-2024-06-14-15-36-40-566/ > output/ > model/
model/
Copy S3 URI
Objects
Properties
Objects (15) Info
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Find objects by prefix
Name
Type
Last modified
Size
Storage class
__script_info__.json
json
June 14, 2024, 11:49:18 (UTC-04:00)
164.0 B
Standard
added_tokens.json
json
June 14, 2024, 11:49:38 (UTC-04:00)
21.0 B
Standard
config.json
json
June 14, 2024, 11:49:25 (UTC-04:00)
705.0 B
Standard
generation_config.json
json
June 14, 2024, 11:49:57 (UTC-04:00)
132.0 B
Standard
inference.py
py
June 14, 2024, 11:49:18 (UTC-04:00)
0 B
Standard
model-00001-of-00003.safetensors
safetensors
June 14, 2024, 11:49:25 (UTC-04:00)
4.6 GB
Standard
model-00002-of-00003.safetensors
safetensors
June 14, 2024, 11:49:04 (UTC-04:00)
4.6 GB
Standard
model-00003-of-00003.safetensors
safetensors
June 14, 2024, 11:49:18 (UTC-04:00)
3.3 GB
Standard
model.safetensors.index.json
json
June 14, 2024, 11:49:38 (UTC-04:00)
23.4 KB
Standard
serving.properties
properties
June 14, 2024, 11:49:25 (UTC-04:00)
205.0 B
Standard
special_tokens_map.json
json
June 14, 2024, 11:49:38 (UTC-04:00)
552.0 B
Standard


aws
Services
Search
[Alt+S]
N. Virginia
voclabs/user3269235=e685f8fa-0da6-11ef-b19f-5f34d7bd81fb @ 72...
Amazon S3 > Buckets > sagemaker-us-east-1-728432338727 > meta-textgeneration-llama-2-7b-2024-06-14-15-36-40-566/ > profiler-output/ > framework/
framework/
Copy S3 URI
Objects
Properties
Objects (1) Info
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more
Find objects by prefix
1
Name
Type
Last modified
Size
Storage class
training_job_end.ts
ts
June 14, 2024, 11:49:40 (UTC-04:00)
0 B
Standard


aws
Services
Search
[Alt+S]
N. Virginia
voclabs/user3269235=e685f8fa-0da6-11ef-b19f-5f34d7bd81fb @ 72...
Amazon S3 > Buckets > sagemaker-us-east-1-728432338727 > meta-textgeneration-llama-2-7b-2024-06-14-15-36-40-566/ > profiler-output/ > system/
system/
Copy S3 URI
Objects
Properties
Objects (2) Info
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more
Find objects by prefix
1
Name
Type
Last modified
Size
Storage class
incremental/
Folder
-
training_job_end.ts
ts
June 14, 2024, 11:49:40 (UTC-04:00)
0 B
Standard


Amazon S3 > Buckets > sagemaker-us-east-1-728432338727 > meta-textgeneration-llama-2-7b-2024-06-14-15-36-40-566/ > profiler-output/ > system/ > incremental/
incremental/
Copy S3 URI
Objects
Properties
Objects (1) Info
Copy S3 URI
Copy URL
Download
Open
Delete
Actions
Create folder
Upload
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more
Find objects by prefix
1
Name
Type
Last modified
Size
Storage class
2024061415/
Folder
-